Summary Report

By *Anitha Ammati* & *Apoorva Srinivas*

July 2021

Summary

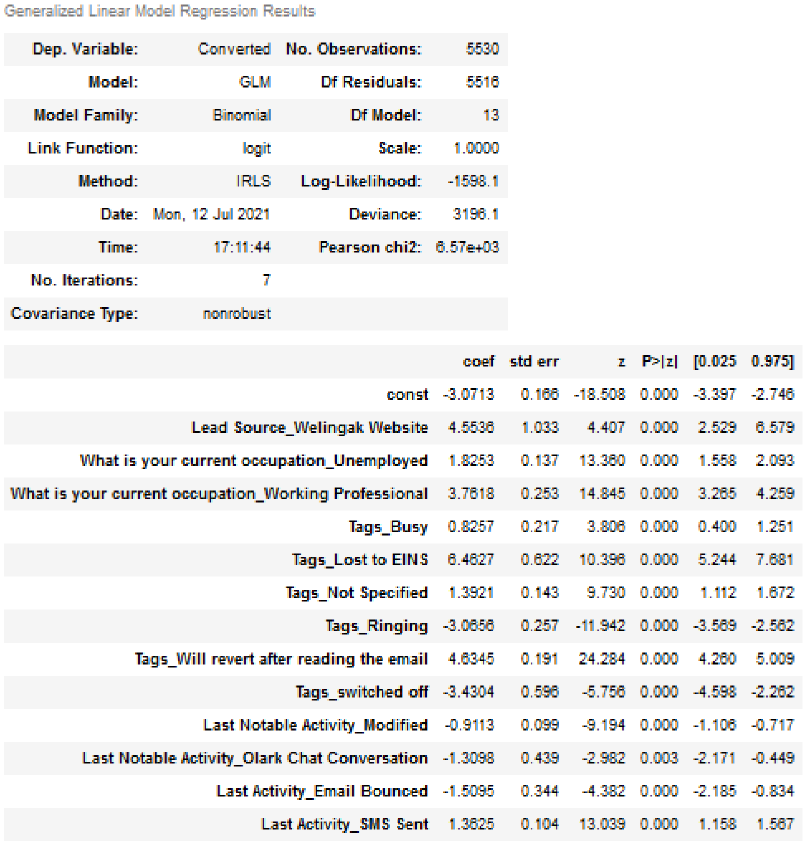
The following steps explain how we proceeded with the Lead Scoring assignment:

1. We downloaded the given dataset and understood the data using the data dictionary.
2. Once we had an idea about the given data, we proceeded with creating a Jupyter notebook and importing the data into a Pandas Dataframe.
3. We realized that there are a lot of missing values and many of the column values had redundant values.
4. Our strategy was to concentrate most of our time on the data preparation phase. So we took time to decide on how to handle/impute the missing values, outliers in the data.
5. We studied each of the variables and individually treated the missing values.
6. During the Exploratory Data Analysis, we conducted the univariate analysis and bivariate analysis of the variables with the target variable. This gave us a lot of insights into the data.
7. In some cases, we combined the values with low count into a single category. This helped in managing the imbalance of data to an extent. However, when the data was highly imbalanced, we dropped the feature.
8. Once the data was clean, we proceeded with the model building process.
9. In this step, we first split the data into train and test data sets.
10. The train set features were scaled and the values were transformed for consistency.
11. Initially, we ran the Generalized Linear Model (GLM) using all the available features. This was not a good model.
12. We used the Recursive Feature Elimination (RFE) for feature selection. We selected the top 15 features.

13. The model ran on these 15 features. The Variance Inflation Factor (VIF) was calculated on these features and all the values were within the limit. But, there were a lot of features which had a p-value close to 1. This indicated that the feature was not significant.

14. We eliminated the feature one by one and ran the model on the remaining features until we got an acceptable p-value of less than 0.05 for all the features.

15. The final model consisted of 13 features. This model had significant p-values and the VIF values were good.



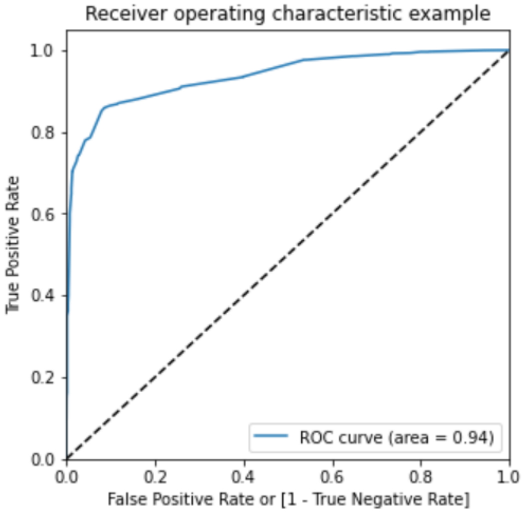
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	5530
Model:	GLM	Df Residuals:	5516
Model Family:	Binomial	Df Model:	13
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1598.1
Date:	Mon, 12 Jul 2021	Deviance:	3196.1
Time:	17:11:44	Pearson chi2:	6.57e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

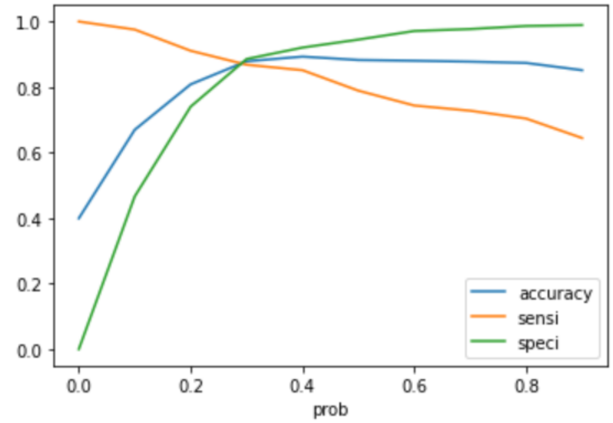
	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.0713	0.166	-18.508	0.000	-3.397	-2.746
Lead Source_Welingak Website	4.5536	1.033	4.407	0.000	2.529	6.579
What is your current occupation_Unemployed	1.8253	0.137	13.360	0.000	1.558	2.093
What is your current occupation_Working Professional	3.7618	0.253	14.845	0.000	3.265	4.259
Tags_Busy	0.8257	0.217	3.806	0.000	0.400	1.251
Tags_Lost to EINS	6.4627	0.622	10.396	0.000	5.244	7.681
Tags_Not Specified	1.3921	0.143	9.730	0.000	1.112	1.672
Tags_Ringing	-3.0656	0.257	-11.942	0.000	-3.569	-2.562
Tags_Will revert after reading the email	4.6345	0.191	24.284	0.000	4.260	5.009
Tags_switched off	-3.4304	0.596	-5.756	0.000	-4.598	-2.262
Last Notable Activity_Modified	-0.9113	0.099	-9.194	0.000	-1.106	-0.717
Last Notable Activity_Olark Chat Conversation	-1.3098	0.439	-2.982	0.003	-2.171	-0.449
Last Activity_Email Bounced	-1.5095	0.344	-4.382	0.000	-2.185	-0.834
Last Activity_SMS Sent	1.3625	0.104	13.039	0.000	1.158	1.567

16. We used the final model to make predictions on the train set and the test set.

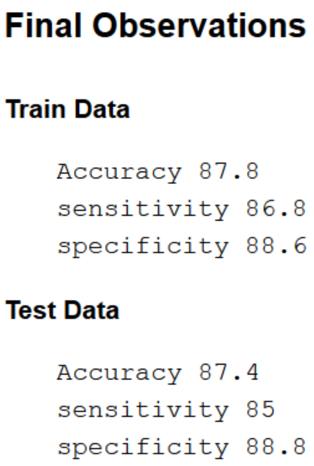
17. The ROC curve was close to the left hand border and the top border. Hence we concluded that the test is accurate.



18. To find the optimal cut-off point, we ran the model against different points and subsequently plotted their accuracy, sensitivity and specificity against the probabilities. We chose the cut-off point as 0.3 as the line graphs intersected at this point.



19. Later, we predicted on the test set using this cut-off point.
20. The evaluation metrics on the test set provided similar values as the train set. This indicates that it is a good model to be considered.



Final Observations

Train Data

```
Accuracy 87.8
sensitivity 86.8
specificity 88.6
```

Test Data

```
Accuracy 87.4
sensitivity 85
specificity 88.8
```

21. Finally, we calculated the lead scores for the test set data.

22. The final model is equal to:

```
p(Converted) = Lead Source_Welingak Website * 4.5536 +
What is your current occupation_Unemployed * 1.8253 +
What is your current occupation_Working Professional *
3.7618 + Tags_Busy * 0.8257 + Tags_Lost to EINS * 6.4627
+ Tags_Not Specified * 1.3921 - Tags_Ringing * 3.0656 +
Tags_Will revert after reading the email * 4.6345 -
Tags_switched off * 3.4304 - Last Notable
Activity_Modified * 0.9113 - Last Notable Activity_Olark
Chat Conversation * 1.3098 - Last Activity_Email Bounced
* 1.5095 + Last Activity_SMS Sent * 1.3625
```

Learnings

- Without handling the low count categories in a column, the imbalanced classes led to a biased model. Hence, we had to reiterate and handle these categories before we built the model.
- The outliers also skewed our analysis in the first run. Hence, we had to handle the outlier values in the ‘Total Visits’ column.
- The univariate analysis of the variables were important for us to understand the data imbalance.

- We built another model with good evaluation metrics and low feature count. But the features consisted of mainly one type of categorical dummy variables (i.e Tags). This would not add much business value since ‘Tags’ variable is an internal variable created by the education company for internal purposes and would not give much variance for the model created. Hence, we decided not to proceed with the model.